Omni-JDocVQA: A Comprehensive Japanese Benchmark for Visual Document Understanding via Situation-Driven QA Generation

Anonymous ACL submission

Abstract

Existing Document Visual Question Answering (DocVQA) benchmarks employ an inverted setup where annotators read documents before creating questions. This causal flaw generates queries that significantly deviate from those observed in real-world usage scenarios. Furthermore, in the Japanese-speaking community, this limitation is compounded by biased document coverage and a lack of document-type labeling, making realistic evaluation even more difficult. To address these challenges, we propose Situation-Driven QA Generation that formulates queries under simulated user scenarios prior to document exposure. Applying this method, we constructed Omni-JDocVQA, a comprehensive Japanese benchmark featuring broad document coverage with a newly designed taxonomy. Our evaluation of 15 Vision-Language Models (VLMs) reveals the generalization performance and specific challenges. We plan to release the benchmark and evaluation code[1].

1 Introduction

With the rapid advancement of Vision Language Models (VLMs), research on Visual Document Understanding (VDU) (Ding et al., 2025), which involves the direct interpretation of document images, has gained significant momentum. At the core of this field lies the task of Document Visual Question Answering (DocVQA) (Mathew et al., 2021b,a; Tanaka et al., 2023; Masry et al., 2022; Landeghem et al., 2023; Ma et al., 2024; Onami et al., 2024; Stockmark Inc., 2025; Ricoh Co., Ltd., 2025). In the English-speaking community, sophisticated benchmarks have been proposed to reflect the diversity and complexity of real-world documents, such as DUDE (Landeghem et al., 2023) for multi-domain coverage and MMLongBench-Doc (Ma et al., 2024) for long-context understanding.

However, existing DocVQA datasets across languages suffer from an inversion of question-document causality in their annotation design. In a standard question generation process, annotators first read a document and then formulate questions based on its visible content. In real-world applications, however, the causal relationship is precisely the opposite: a user possesses a pre-existing question rooted in a specific context or information need and subsequently explores documents to resolve it. In this conventional process, questions implicitly and unconsciously incorporate textual or visual features (layout clues) from the target page. Consequently, this approach creates QA tasks composed of unrealistic queries far removed from actual use cases, making it difficult to accurately evaluate realistic reasoning capabilities.

In addition to this methodological limitation, existing Japanese VDU benchmarks (Onami et al., 2024; Stockmark Inc., 2025; Ricoh Co., Ltd., 2025) remain underdeveloped compared to their English counterparts, suffering from biased document coverage and a lack of detailed document-type labeling. These two limitations in the evaluation documents prevent a detailed analysis of diverse and real-world documents.

To address these challenges simultaneously, we propose a new paradigm in dataset construction: the **Situation-Driven QA Generation**. Instead of initiating the process from document reading, our framework first simulates and generates a document-independent *user scenario*, constructing clear and specific questions without any prior knowledge of the target document's detailed content. These context-driven queries are then mapped back to a massive and diverse document collection using dense vector retrieval to re-pair them with candidate grounding pages. This pipeline com-

[1]The evaluation code will be released under the Apache 2.0 license. This benchmark is available only for purposes that comply with Article 30-4 of the Copyright Act in Japan.

pletely eliminates the conventional inversion of causality, establishing a rigorous evaluation foundation for measuring true document comprehension capabilities.

We constructed **Omni-JDocVQA** as a novel Japanese VDU benchmark to comprehensively realize this methodology. Evaluating Japanese PDF files extracted from Common Crawl, our benchmark adopts a comprehensive dual-taxonomy spanning 9 domains and 9 formats. Crucially, Omni-JDocVQA incorporates unanswerable ("No Evidence") cases, which are annotated with explicit, human-authored detailed reasons for unanswerability in a free-text format. This prevents models from fabricating answers based on prior knowledge (i.e., hallucinating) and serves as a powerful stress test to strictly verify visual grounding. Through extensive experiments evaluating 15 publicly available Japanese and multilingual VLMs, we expose severe processing bottlenecks in current models.

Our primary contributions are summarized as follows. First, we designed realistic QA tasks by proposing a situation-driven pipeline that mirrors real-world search behaviors. Second, to enable detailed analysis and verification of visual grounding, we introduced a domain-and-format dual-taxonomy based on document types and explicitly annotated human-authored reasons for unanswerable cases. Finally, through benchmarking 15 models, we identified critical performance bottlenecks, revealing that while frontier multilingual models dominate, localized Japanese models suffer from severe hallucinations and output collapse on multi-page documents.

2 Related Work

2.1 Document Visual Question Answering

Document Visual Question Answering (DocVQA) has served as a benchmark task for evaluating Visual Document Understanding (VDU) capabilities, requiring models to simultaneously process multimodal information such as textual content, visual appearance, and structural layouts. Early benchmarks in this domain primarily focused on single-page document images. For example, the pioneering DocVQA dataset (Mathew et al., 2021b) introduced a diverse collection of reports and forms to evaluate text-based QA on rich layouts. Subsequent research extended this line of work to element-rich documents in InfographicVQA (Mathew et al., 2021a) and specific charts in ChartQA (Masry et al., 2022).

Driven by the need to evaluate Vision-Language Models (VLMs) under more realistic and complex scenarios, recent benchmarks have shifted their focus toward multi-page and multi-domain settings. The DUDE benchmark (Landeghem et al., 2023) expanded the evaluation horizon to multi-page multi-domain document comprehension, requiring models to navigate long-context visual layouts. More recently, MMLongBench-Doc (Ma et al., 2024) introduced a rigorous framework to test long-context document understanding across extensive page segments, exposing critical context-window limitations in modern VLMs.

While VDU has advanced, existing datasets suffer from the question-document causality inversion, where annotators thoroughly read documents before creating questions. This conventional pipeline introduces layout clues and textual shortcuts into queries, allowing models to bypass holistic document comprehension. Consequently, these benchmarks fail to strictly evaluate a model's true open-domain search and visual grounding capabilities under realistic scenarios.

Furthermore, existing benchmarks simplistically treat unanswerable questions with rigid negative labels. This oversimplified approach makes it impossible to discern whether a model genuinely understands document layouts. This limitation motivates us to construct a comprehensive benchmark that combines fine-grained multi-labeling with explicit human-authored reasons for unanswerability.

2.2 Non-English and Japanese Document Benchmarks

In addition to the issues described in 2.1, Japanese document benchmarks are heavily constrained by their restricted information sources and document types. For example, JDocQA (Onami et al., 2024) primarily comprises official government publications, while BusinessSlideVQA (Stockmark Inc., 2025) focuses exclusively on business presentation slides. Similarly, JGraphQA (Ricoh Co., Ltd., 2025) specifically targets charts and graphs. While these benchmarks are effective within their specialized scopes, they fail to encompass the extensive variety of document types found in real-world scenarios. Furthermore, the lack of fine-grained document labeling forces them to report monolithic evaluation scores, precluding granular error analysis.

3 Situation-Driven QA Generation

We propose situation-driven QA generation to eliminate the inversion of question-document causality inherent in conventional datasets. Unlike traditional pipelines where annotators read documents before formulating questions, this framework initiates the process by simulating a document-independent user scenario. Proposed pipeline consists of four phases.

Scenario Generation First, a model generates an initial *summary* and *viewing purpose* from a target document, representing an overview of the document and the context in which it is viewed, respectively. Furthermore, to remove direct dependencies on the specific document content, a model subsequently transforms these initial outputs into a document-independent user scenario.

Question Generation Second, a model generates clear and specific questions based on the simulated context. The questions must be clear and specific, and they must be unanswerable without reading the provided document[2].

Question-Document Re-pairing Third, the pipeline maps these isolated queries back to relevant candidates within the document collection using dense retrieval. This causal design effectively replicates realistic open-domain search behaviors, ensuring that the benchmark evaluates true document comprehension without relying on superficial layout clues.

Ground Truth Answer Annotation Finally, human annotators construct ground-truth answers for the resulting question-document pairs based strictly on objective facts within the documents. When a document lacks sufficient evidence to resolve a query, annotators explicitly document the detailed, free-text reasons for unanswerability.

4 Omni-JDocVQA

In this section, we describe the construction procedure of the benchmark tasks and the evaluation methodology for each task.

4.1 Document Collection

The objective of this benchmark is to encompass the diversity of documents found in the real world. To achieve this, we used CCpdf (Turski et al., 2023), a large-scale collection of PDF files extracted from Common Crawl, as our starting point. From this corpus, we non-selectively extracted Japanese PDF files to build our target document set[3]. We performed OCR using Qwen3-VL[4] (Qwen Team, 2025b) to extract text and identified Japanese documents using fasttext-langdetect[5] (Joulin et al., 2016a,b), retaining files with a Japanese language probability exceeding 0.9. Consequently, 70,347 files were identified as valid Japanese PDFs.

4.2 Document Taxonomy and Labeling

We define document types using a dual-taxonomy comprising nine categories each for **Domain** (e.g., *Academic*, *Medical*) and **Format** (e.g., *Reports*, *Manuals*). Each document is labeled according to these categories, which enables not only an overall performance assessment but also a granular analysis restricted to specific classifications. We adopt a multi-label format, allowing a single document to have multiple labels, while assigning an "Other" label to documents that do not fit any predefined category. We employed Qwen3-VL for the labeling process. Detailed label definitions are provided in Table 3 in Appendix A.1.

4.3 QA Annotation Pipeline

Document Sampling To select documents for the QA task, we perform sampling as follows: First, we vectorize the first three pages of each document using the ruri-v3-310m (Tsukagoshi and Sasano, 2024) embedding model. Next, we apply k-means clustering to these vector representations and extract the documents closest to each cluster's centroid. This strategy ensures the diversity of documents within each label while sampling those with representative structures.

Situation-Driven QA Generation We construct QA pairs based on the sampled documents. For the QA construction methodology, we adopt the Situation-Driven QA Generation framework described in Section 3. Below, we provide a detailed explanation of each step, including the specific models utilized.

[2] Details regarding the question generation prompt are provided in Appendix A.2.

[3] PDF files that were unreadable due to encryption or data corruption were excluded in advance.

[4] https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct

[5] https://pypi.org/project/fasttext-langdetect/

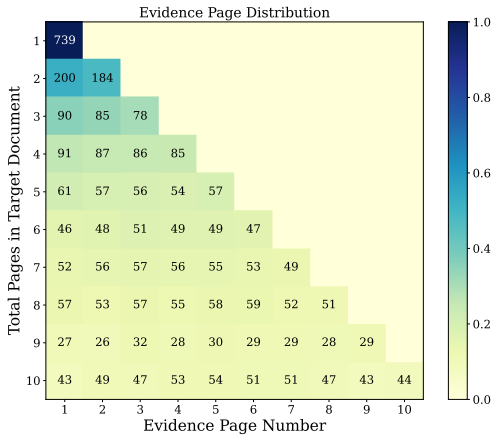

Figure 1: Distribution of evidence within documents across different page counts.

For Scenario Generation, we employed Qwen3-VL to generate initial scenarios from documents, and Qwen3 (Qwen Team, 2025a) to derive similar scenarios from them. For Question Generation, we utilized gpt-oss-120b (OpenAI, 2025) to generate three questions per scenario. For Question-Document Re-pairing, we used ruri-v3-310m for dense retrieval, linking each question with the top-3 retrieved documents as relevant documents. For Ground Truth Answer Annotation, considering the workload of the human annotators, we restricted the target documents to those with 10 pages or fewer. Finally, the QA set constructed through this process consists of 1,791 samples[6].

4.4 Evaluation Methodology

We evaluate the accuracy of model responses using the **LLM-as-a-Judge** framework (Zheng et al., 2023). In this evaluation, the evaluator model is presented with a triplet consisting of the question, the ground-truth answer, and the model-generated response. The judge then determines the degree of alignment between the generated response and the ground truth within the context of the question, using a five-point Likert scale[7]. As Omni-JDocVQA includes unanswerable cases, the evaluated models are also required to generate the reason for unanswerability when an answer cannot be produced.

4.5 Dataset Statistics

Figure 2 illustrates the distribution of samples across document labels, where n denotes the sample size for each category. We ensured that each label, excluding the "Other" category, contains at least approximately 100 cases to maintain statistical significance. Within the final set, 250 cases (representing 14% of the sampled subset) were identified as "No Evidence" cases.

The heatmap in Figure 1 displays the distribution of pages containing the evidence for answers. By analyzing the frequency of evidence pages (horizontal axis) relative to the total number of pages in each document (vertical axis), we observed that the grounding evidence is distributed almost uniformly across the documents rather than being clustered on specific pages. The mean variance of the evidence page position across all documents was 3.03. The balanced distribution of grounding evidence across pages prevents the evaluated models from answering by merely referencing localized features, suggesting that the benchmark evaluates whether they accurately comprehend the entire document.

5 Experiments

5.1 Experimental Setup

Using the Omni-JDocVQA benchmark, we evaluated a suite of Japanese and multilingual Vision Language Models (VLMs) released between 2023 and 2025. All models were obtained from the HuggingFace Hub[8]. During inference, we set the temperature to 0.7 and the maximum number of generation tokens to the maximum context window supported by each model (based on their position embeddings)[9]. To account for stochasticity in model outputs, we reported the average scores over three evaluation runs using different random seeds.

We used gpt-5.1-2025-11-13 as the evaluator model for LLM-as-a-Judge. To ensure deterministic and consistent judgments, we set both the temperature and the seed for the evaluator model to 0.

5.2 Results

Figure 2 shows the evaluation results. The evaluation results broken down by document domain and format provide deeper insights into the structural and contextual processing capabilities of each model, which cannot be captured by the overall scores alone.

[6] We paid annotators 1,100 JPY (approximately 6.90 USD, exchange rate as of May 25, 2026) per annotation.

[7] Details regarding the evaluation prompt are provided in Appendix A.2.

[8] https://huggingface.co/models

[9] Model inference was completed in about 7 days using 8 NVIDIA A100 40GB GPUs.

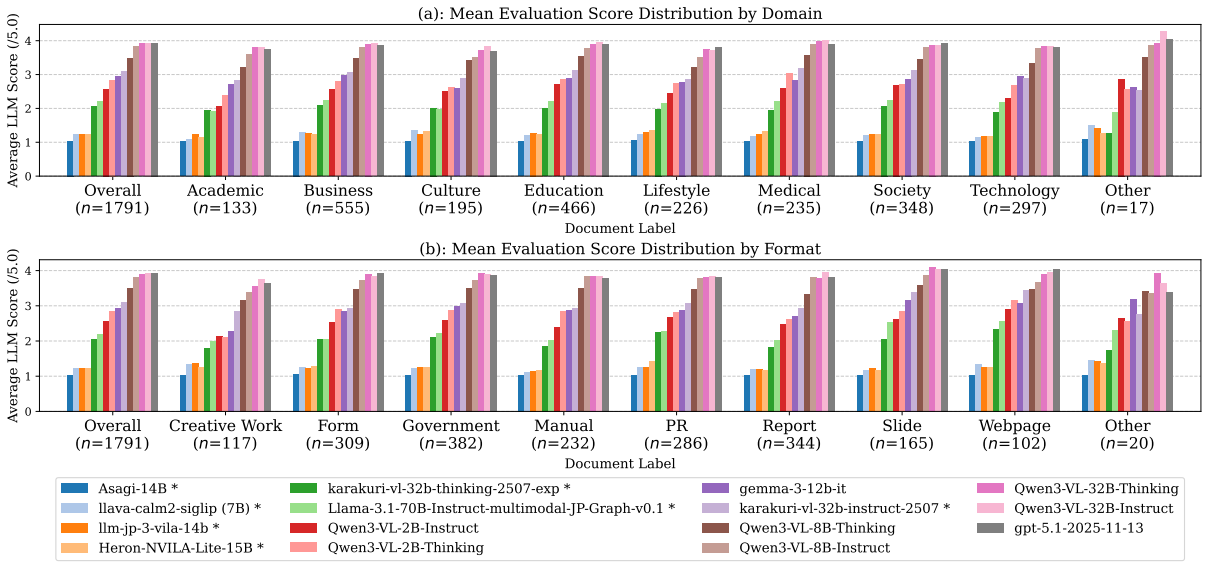

Figure 2: Distribution of mean evaluation scores across document labels. n denotes the number of target documents evaluated in the QA task. Asterisks (*) next to model names in the legend indicate Japanese-specialized models.

Performance Discrepancy between Multilingual and Japanese Models First, a significant performance discrepancy is observed between state-of-the-art multilingual models and models developed or fine-tuned specifically for the Japanese language. Global frontier models, such as gpt-5.1-2025-11-13 and the Qwen3-VL series (e.g., 32B-Instruct), consistently dominate this benchmark, achieving a top overall score of up to 3.93.

In contrast, most Japanese models, with the exception of the karakuri-vl series, remain in the 1-to-2 point range. This indicates that there is still substantial room for improvement in the visual document understanding capabilities of Japanese-centric models.

Advantages of Category-wise Evaluation The primary advantage of evaluating performance at the document-label level is the ability to diagnose a model's unique strengths and weaknesses across specific visual structures and knowledge domains. For instance, observing the scores of karakuri-vl-32b-instruct at the category level reveals that while the model achieves high performance in domains such as *Medical* (3.19) and *Society* (3.12), it remains at relatively lower scores in domains like *Academic* (2.84) and *Lifestyle* (2.85). Similarly, from the perspective of document formats, a cross-model comparison reveals that gpt-5.1—the top-performing model overall—falls behind the runner-up, Qwen-VL-32B-Instruct, in *Creative work*. Consequently, this benchmark makes it possible to comprehensively compare and analyze the detailed characteristics of various models.

6 Analysis

6.1 Analysis of Unanswerable Questions

We investigate the specific behaviors of modern vision-language models (VLMs) when facing unanswerable questions (i.e., queries whose answers are missing from the document), as exemplified in Table 1.

In the first case study, we observe a clear disparity in the quality of refusal when models encounter unanswerable queries. Qwen3-VL-8B-Instruct successfully provides a polite and correct refusal by explicitly explaining the reason why the required information is missing from the document. In contrast, llm-jp-3-vila-14b (Sasagawa et al., 2025) merely responds with a brief negation, "None." ("ありません。"), making it unclear whether the model actually understands the document layout or is simply outputting a generic phrase. Therefore, accurately assessing a model's ability to correctly identify unanswerable cases is crucial for real-world applications.

The second case study highlights a more severe failure mode: the response from llm-jp-3-vila-14b contains content entirely unrelated to the provided document image. This result strongly suggests that an unanswerable query can act as a direct catalyst that triggers critical hallucinations in certain VLMs. When operating under the implicit assumption that

Question	Gold Answer	Model Response → Evaluation Score
What are the evaluation items common to both reports? (2つの報告書で評価項目として共通している項目は何ですか？)	Only one document is provided, and it is a self-evaluation sheet for an exam, not a report; thus, it cannot be answered. (Qwen: 提示された文書は1つのみで、また報告書ではなく検定試験の自己評価シートであるため解答できません。)	**Qwen**: The document only describes one report, which is the "self-evaluation sheet for an exam." Since the "two reports" mentioned in the question do not exist, it is impossible to identify any common items. → **5.0** (Qwen: 文書には「検定試験の自己評価シート」という1つの報告書のみが記載されています。質問で言及されている「2つの報告書」は文書内に存在しないため、評価項目の共通点を特定することはできません。) → **5.0** **LLM-jp**: None. → **1.0** (LLM-jp: ありません。) → **1.0**
What is the purpose of this study? (この研究の目的は何ですか？)	The provided document is intended for proposing content-matched advertisements; thus, it cannot be answered. (提示された文書は、コンテンツマッチ広告を提案するための文書であるため解答できません。)	**LLM-jp**: The purpose of this study is to show how the human brain perceives sounds and images. → **1.0** (LLM-jp: この研究の目的は、人間の脳がどのように音やイメージを知覚するかを示すことです。) → **1.0**

Table 1: Examples of Unanswerable Questions (Qwen: Qwen3-VL-8B-Instruct, LLM-jp: llm-jp-3-vila-14b).Each case displays an English translation above the original Japanese sentences.

every question has an answer, models often neglect visual evidence and fabricate false text based on language priors. Therefore, our unanswerable questions serve as a challenging stress test to evaluate whether they can recognize their limitations and properly process visual information.

6.2 Performance on Multi-page Documents

We analyze the performance of each VLM on multi-page documents within our benchmark. In real-world scenarios, documents such as business reports and websites often span multiple pages, making multi-page comprehension indispensable.

Our experiments reveal that as the number of input pages increases, Japanese VLMs frequently generate broken responses, such as a random series of symbols or repetitive words. This failure is likely because the sharp increase in input images produces a massive number of visual tokens, resulting in decrease in processing capacity of these models (Wang et al., 2025).

To investigate this issue further, we conducted additional evaluations using other Japanese VLMs, sarashina2.2-vision-3b and Stockmark-2-VL-100B-beta[10]. We found that as the page count grows, these models quickly hit their maximum input token limits (context window limitations). This leads to processing crashes, making it completely impossible to calculate evaluation scores. This phenomenon exposes a fundamental bottleneck in current Japanese VLMs: even before considering output quality, these models lack the structural capacity to accept large amounts of visual data simultaneously.

On the other hand, Qwen3-VL series did not exhibit any output collapse or processing errors, maintaining stable response generation. This resilience is attributed to the fact that the model was specifically optimized to handle extensive document images spanning dozens of pages (Qwen Team, 2025b). These findings demonstrate that current Japanese VLMs still face significant barriers in their ability to process massive amounts of visual data concurrently.

6.3 Reliability of Automated Evaluation

To verify the reliability of the LLM-as-a-judge framework adopted in this benchmark, we conducted a human evaluation of the model responses. We selected a total of five models, including multilingual models (gpt5.1-2025-11-13, gemma-3-12b-it, and Qwen-VL-32B-Instruct) and Japanese models (karakuri-vl-32b-instruct and llm-jp-3-vila-14b). We randomly sampled 20 instances from each model, yielding a total of 100 triplets of question, reference answer, and model response for evaluation. Each instance was evaluated by two independent human annotators based on the 5-point Likert scale defined in Section 4.4, and the final human

[10] https://huggingface.co/sbintuitions/sarashina2.2-vision-3b, https://huggingface.co/stockmark/Stockmark-2-VL-100B-beta

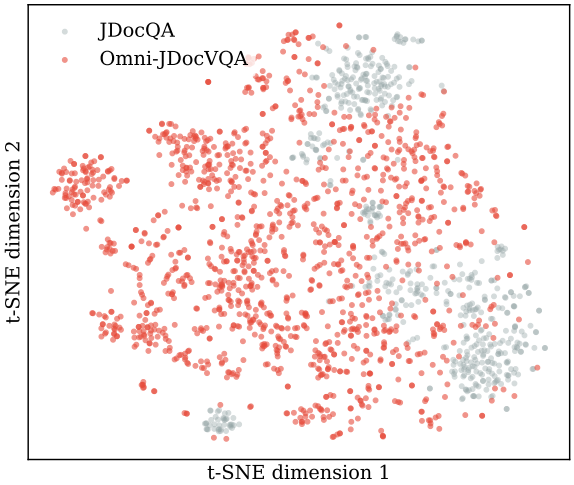

Figure 3: t-SNE visualization of document image embeddings. The light grey background points represent the overall distribution of the JDocQA dataset, while the red dots highlight the distribution of the Omni-JDocVQA dataset for each respective category.

score for each response was determined by taking the average of their scores[11].

The Pearson correlation coefficient between the automated evaluation and the averaged human evaluation was calculated to be $r = 0.90$ (95% CI : $[0.856, 0.932]$), indicating a strong positive correlation. This correlation is statistically highly significant ($p < 0.05$), demonstrating that the automated evaluation closely mimics human judgment.

6.4 Diversity and Taxonomy of Document Corpus

Existing Japanese benchmarks focus primarily on official government gazettes and business documents, which introduces a challenge regarding limited document diversity. To evaluate the diversity of our target documents, we projected the document image embeddings into a two-dimensional space using t-SNE (van der Maaten and Hinton, 2008) for qualitative analysis. We utilized Qwen3-VL-Embedding-8B[12] (Li et al., 2026) to extract visual embeddings from the first page of each document image.

Figure 3 displays the visualization results for the entire corpus. The distribution of JDocQA, the existing benchmark, is heavily concentrated within a few narrow regions characterized by specific visual features. In sharp contrast, the proposed Omni-JDocVQA widely disperses across the entire space while completely covering the regions occupied by JDocQA. This clear expansion demonstrates that Omni-JDocVQA is not constrained to specific document categories but is composed of highly diverse document types.

In addition to the overall dataset distribution, we visualized the embeddings categorized by specific document types. As shown in the t-SNE plots in Figure 4, JDocQA (grey) is restricted to a few high-density clusters resulting from specific visual layouts (such as government documents or uniform presentation slides). On the contrary, the proposed Omni-JDocVQA (red) thoroughly blankets the feature space. Notably, for categories such as "Input Form" and "Manual" in the format labels, and "Technology" and "Healthcare" in the domain labels, distinct distributions are formed in unmapped regions where the existing benchmark contains no data points. These findings visually confirm that Omni-JDocVQA extensively covers a broad spectrum of real-world document variations.

To quantitatively evaluate document diversity, we calculated the Average Pairwise Distance (APD) and Mean Distance to Centroid (MDC) in the embedding space. To ensure a strictly fair comparison regardless of sample size differences, we randomly sampled $N = 500$ documents from each dataset over 100 independent iterations and computed their average values. Under identical sample size conditions, Omni-JDocVQA achieved a 9.4% higher APD (91.58 vs. 83.68) and a 9.3% higher MDC (64.71 vs. 59.21) compared to JDocQA. These empirical results mathematically substantiate that our corpus possesses inherently high diversity across the unique visual and structural features of individual documents.

6.5 Realism Asymmetry in Question Design

To verify whether the questions included in our dataset are realistic in actual usage scenarios, we conducted a human evaluation. We randomly sampled 100 questions each from Omni-JDocVQA and JDocQA, yielding a total evaluation set of 200 questions. In this evaluation, independent annotators scored each question on a 3-point scale (1–3)[13] based on the following three distinct criteria. Under this setup, higher scores indicate greater question realism:

[11] We paid annotators 1,650 JPY (approximately 10.38 USD, exchange rate as of May 25, 2026) per annotation.

[12] https://huggingface.co/Qwen/Qwen3-VL-Embedding-8B

[13] We paid annotators 1,650 JPY (approximately 10.38 USD, exchange rate as of May 25, 2026) per annotation.

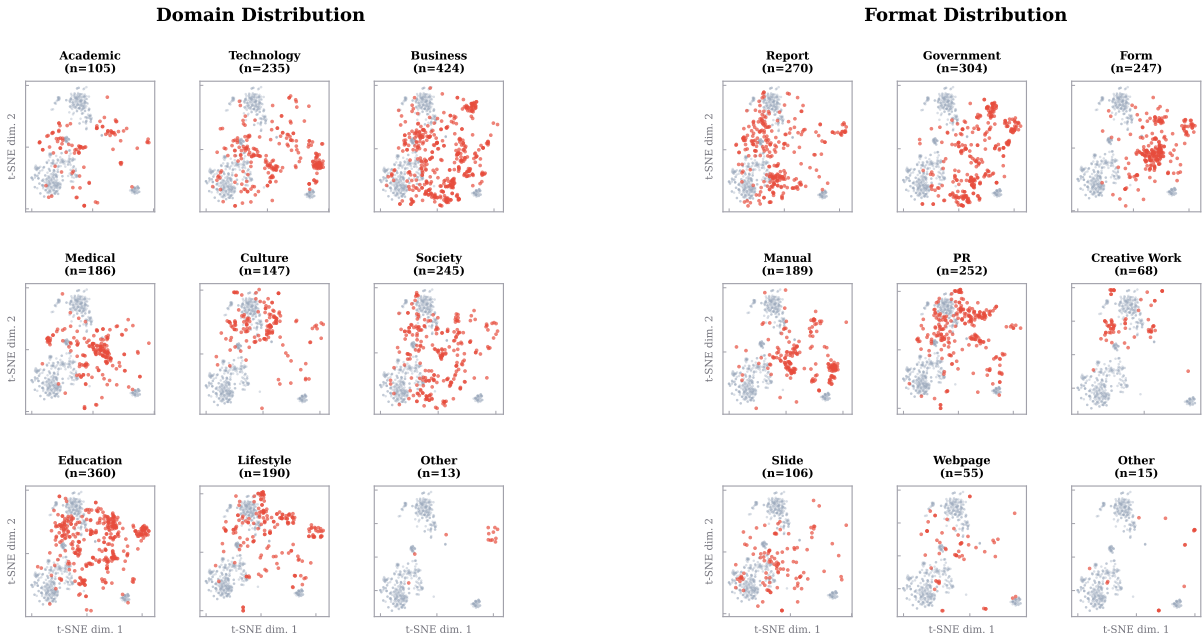

Figure 4: t-SNE visualization of document image embeddings across different domains (left) and formats (right). The light grey background points represent the overall distribution of the JDocQA dataset, while the red dots highlight the distribution of the Omni-JDocVQA dataset for each respective category.

Dataset	Doc. Unfamiliarity	Purpose Specificity	Question Naturalness
JDocQA	1.830	2.220	2.195
Omni-JDocVQA (Ours)	**2.040**	**2.395**	**2.425**

Table 2: Comparison of question quality scores between JDocQA and Omni-JDocVQA. Scores are based on a human evaluation utilizing a 3-point scale (1–3), where higher scores indicate greater question realism. **Bold** scores indicate higher scores compared to JDocQA.

- **Document Unfamiliarity:** Whether it is perceptible that the questioner is asking without prior knowledge of the document's content (i.e., simulating a user who genuinely does not know the answer).
- **Purpose Specificity:** Whether a realistic user action or objective can be clearly imagined from the text of the question.
- **Question Naturalness:** Whether the phrasing feels natural and seamless as an input to an actual QA system.

We adopted the existing Japanese document VQA benchmark, JDocQA, as our baseline.

As shown in the Table 2, Omni-JDocVQA outperforms JDocQA across all three evaluation criteria. This outcome demonstrates that our proposed situation-driven question generation process functions highly effectively. Furthermore, statistical tests confirm that these improvements are statistically significant ($p < 0.05$). These results validate that Omni-JDocVQA enables a more realistic evaluation of actual QA scenarios.

7 Conclusion

In this study, we proposed Omni-JDocVQA, a benchmark designed to evaluate Japanese Visual Document Understanding capabilities. To avoid the bias that arises when questions are generated after inspecting the document, we introduced Situation-Driven QA Generation, which first defines realistic user situations and then generates questions based on them. Based on this method, we constructed Omni-JDocVQA with 1,363 documents and 1,791 QA pairs spanning various domains and formats. Through human evaluation, we showed that Omni-JDocVQA provides more realistic and natural QA pairs than existing benchmarks. Our benchmark enables fine-grained performance analysis across diverse document types using realistic questions.

Limitations

While Omni-JDocVQA provides a robust evaluation framework, several limitations should be acknowledged. First, to manage the human annotation workload, the target documents were restricted to a maximum of 10 pages. Although this length is sufficient to expose severe context-window and visual-token bottlenecks in localized models, it does not fully encompass extremely long documents such as comprehensive corporate reports or product manuals spanning dozens of pages. Second, although the final ground-truth answers and evidence pages are strictly verified by human annotators, the initial context and question generation rely on automated models which may introduce implicit stylistic patterns or structural biases into the pipeline. However, our human evaluation confirmed that the questions generated through this approach achieve significantly higher realism in actual usage scenarios compared to the existing benchmark, JDocQA. Third, we did not conduct an ablation study using questions generated without Situation-Driven QA Generation. However, we compared the constructed benchmark with existing benchmarks and evaluated whether its questions are more natural and realistic. Although this is not a strict comparison, it provides suggestive evidence for the effectiveness of the proposed method.

Ethical Considerations

All annotators in this work received appropriate compensation based on the agreement between the contracted annotation vendors and the researchers. The cost estimation provided by the vendor was carefully calculated to ensure that the hourly wage for each annotator did not fall below the statutory minimum wage in Japan.

References

Yihao Ding, Siwen Luo, Yue Dai, Yanbei Jiang, Zechuan Li, Geoffrey Martin, and Yifan Peng. 2025. A Survey on MLLM-based Visually Rich Document Understanding: Methods, Challenges, and Emerging Trends. *Preprint*, arXiv:2507.09861.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759*.

Jordy Van Landeghem, Rubèn Pérez Tito, Łukasz Borchmann, Michal Pietruszka, Pawel J'oziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Ackaert, Ernest Valveny, Matthew B. Blaschko, Sien Moens, and Tomasz Stanislawek. 2023. Document Understanding Dataset and Evaluation (DUDE). *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19471–19483.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. InfographicVQA. *arXiv:2104.12756*.

Minesh Mathew, Dimosthenis Karatzas, and C.V. Jawahar. 2021b. DocVQA: A Dataset for VQA on Document Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209.

Eri Onami, Shuhei Kurita, Taiki Miyanishi, and Taro Watanabe. 2024. JDocQA: Japanese Document Question Answering Dataset for Generative Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9503–9514.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card.

Qwen Team. 2025a. Qwen3 Technical Report.

Qwen Team. 2025b. Qwen3-Vl Technical Report.

Ricoh Co., Ltd. 2025. JGraphQA. https://huggingface.co/datasets/r-g2-2024/JGraphQA.

Keito Sasagawa, Koki Maeda, Issa Sugiura, Shuhei Kurita, Naoaki Okazaki, and Daisuke Kawahara. 2025. Constructing Multimodal Datasets from Scratch for Rapid Development of a Japanese Visual Language Model. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 470–484.

Stockmark Inc. 2025. Business SlideVQA. https://github.com/stockmarkteam/business-slide-questions.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images. In *AAAI*.

Hayato Tsukagoshi and Ryohei Sasano. 2024. Ruri: Japanese General Text Embeddings. *Preprint*, arXiv:2409.07737.

Michał Turski, Tomasz Stanisławek, Karol Kaczmarek, Paweł Dyda, and Filip Graliński. 2023. CCpdf: Building a High Quality Corpus for Visually Rich Documents from Web Crawl Data. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part III*, page 348–365.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using T-Sne. *JMLR*, 9(86):2579–2605.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, Yangqiu Song, and Mark Steedman. 2025. MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly. In *The 39th (2025) Annual Conference on Neural Information Processing Systems*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge With MT-bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

A Appendix

A.1 Document Labels

The definitions of the document labels are shown in Table 3.

A.2 Prompt Templates

The prompts used in this study are shown below.

Template prompts of text extraction

Please transcribe the text, tables, and mathematical expressions in the following document image into Markdown format without any modifications.
Do not generate anything other than the transcribed content.

Template prompts of document labeling (Domain)

You are an annotator classifying document types.
Please select at least one appropriate label (domain) for the following document.

domain (represents the content field described in the document)
- Academic: Academic research and specialized knowledge.
- Technology: Application of scientific knowledge and product development.
- Business: Corporate activities, markets, finance, etc.
- Medical: Medicine, health, treatment, beauty, etc.
- Culture: Creative works such as art, literature, and history.
- Society: Public systems and social issues.
- Education: Learning activities and educational systems.
- Lifestyle: Daily life, hobbies, and consumer activities.
- Other: Documents that do not fall under the above labels.

Conditions:
- You must select at least one label.
- Do not output anything other than the label(s).

Domain		Format	
Label	Description	Label	Description
Academic	Academic research and specialized knowledge.	Report	Reports on the results of investigations, analyses, or activities.
Technology	Application of scientific knowledge and product development.	Government	Documents issued by public institutions with legal binding or procedural roles.
Business	Corporate activities, markets, finance, etc.	Form	Documents designed for filling out or selecting information.
Medical	Medicine, health, treatment, beauty, etc.	Manual	Explanations of usage, procedures, and guidelines.
Culture	Creative works such as art, literature, and history.	PR	Notifications or advertisements of products, services, events, etc.
Society	Public systems and social issues.	Creative Work	Delivery of stories, knowledge, and artistic expressions.
Education	Learning activities and educational systems.	Slide	Supplementary materials for oral presentations.
Lifestyle	Daily life, hobbies, and consumer activities.	Webpage	Documents created specifically to be published on the Web.
Other	Documents that do not fall under the above labels.	Other	Documents that do not fall under the above labels.

Table 3: Definitions of document labels

Template prompts of document labeling (Format)

You are an annotator classifying document types.
Please select at least one appropriate label (format) for the following document.

format (represents the structural form or purpose of the document)
- Report: Reports on the results of investigations, analyses, or activities.
- Government: Documents issued by public institutions with legal binding or procedural roles.
- Form: Documents designed for filling out or selecting information.
- Manual: Explanations of usage, procedures, and guidelines.
- PR: Notifications or advertisements of products, services, events, etc.
- Creative Work: Delivery of stories, knowledge, and artistic expressions.
- Slide: Supplementary materials for oral presentations.
- Webpage: Documents created specifically to be published on the Web.
- Other: Documents that do not fall under the above labels.

Conditions:
- You must select at least one label.
- Do not output anything other than the label(s).

Template prompts of question generation

You are an annotator whose task is to create questions designed to obtain the necessary answers from the content of a document.
Please create three questions based on the provided situation for referencing the document.
However, you must strictly adhere to the following conditions:

1. Ensure the questions are clear and specific in content.
2. Make sure the questions yield the same answer regardless of who asks them or when.
3. The questions and answers must not require specialized knowledge.
4. Create questions under the assumption that the answer can be obtained from the document.
5. Create questions that seek objective facts.
6. Create questions that cannot be answered without reading the document.
7. Avoid using proper nouns.
8. Write the questions in fluent and natural Japanese.
9. If you cannot create questions that satisfy the above conditions from the given situation, reply only with 'None'.

Template prompts of scoring evaluation (1/2)

Please evaluate the response generated by a model performing question answering on a document.

Evaluation Targets:
You will be provided with the following triplet as the target of evaluation:
- Question: A question regarding the content of the document.
- Gold Answer: An accurate, necessary, and sufficient answer based on the document.
- Prediction: The response generated by the model being evaluated.

Response Conditions:
The response to be evaluated was generated under the following conditions:
1. The response must be generated based solely on objective facts written in the document.
2. The response must be generated in Japanese text.
3. If an answer cannot be generated from the content of the document, the response must state the reason for unanswerability.

Template prompts of scoring evaluation (2/2)

Scoring Criteria:
Based on how semantically aligned the response is with the gold answer within the context of the question, score the evaluation on the following 5-point scale:
5: Excellent - Perfectly matches the gold answer.
4: Good - There are minor differences between the gold answer and the response.
3: Fair - There are clear differences between the gold answer and the response.
2: Poor - The gold answer and the response are significantly different.
1: Very Poor - The gold answer and the response are completely different, or the response is not written primarily in Japanese, or the response fails to explicitly state the reason for unanswerability when a reason is required as the correct answer.

Output Format:
- Output only the score (an integer from 1 to 5).
- Do not include any explanations or extra text.

Question: Question
Gold Answer: Answer
Prediction: Prediction
Score:

Template prompts of text extraction (original)

以下の文書画像のテキスト、表、数式を改変せずにMarkdown形式で転記してください。
転記した内容以外は生成しないでください。

Template prompts of document labeling (Domain; original)

あなたは文書の種類を分類するアノテーターです。
次に示す文書に適切なラベル（domain）を、必ず1つ以上選んでください。

domain（文書に記載されている内容分野を表す）
- 学術：学問研究や専門知識に関する分野
- 技術：科学的知識の応用や製品・システム開発に関する分野
- ビジネス：企業活動、市場、金融、商業に関する分野
- 医療：医学、健康、治療、美容に関する分野
- 文化：芸術、文学、歴史、宗教など人間の創作物に関する分野
- 社会：社会のルール、公共システム、社会全体の課題に関する分野
- 教育：学習・教育活動や教育制度に関する分野
- 生活：日常生活、趣味、家庭、消費活動に関する分野
- その他：上記ラベルに該当しない分野

条件：
- ラベルは必ず1つ以上選んでください。
- ラベル以外は出力しないでください。

Template prompts of document labeling (Format; original)

あなたは文書の種類を分類するアノテーターです。
次に示す文書に適切なラベル（format）を、必ず1つ以上選んでください。

format（文書の形式や目的を表す）
- レポート：調査、分析、活動の結果を報告する文書
- 公的文書：公的機関が発行する法的拘束や手続きを目的とする文書
- 入力フォーム：情報を記入・選択することを目的とした文書
- マニュアル：使い方や手順、利用方法を説明するための文書
- 広報：商品、サービス、イベントなどを広く知らせるための文書
- 創作物：物語や知識の提供、芸術表現を目的とした文書
- スライド：口頭発表やプレゼンテーションの補助を目的とした文書
- ウェブサイト：ウェブ上に公開されることを前提とした文書
- その他：上記ラベルに該当しない文書

条件：
- ラベルは必ず1つ以上選んでください。
- ラベル以外は出力しないでください。

Template prompts of question generation (original)

あなたは文書に書かれた内容から必要な回答から得ることを目的とした質問を作成するアノテーターです。
これから提供する文書を参照する状況に基づいて、質問を3件作成してください。
ただし、以下の条件に従ってください。

1. 質問は明確で具体的な内容で構成してください
2. 誰がいつ質問しても同じ回答が得られる質問にしてください
3. 質問および回答には、専門知識を必要としない内容にしてください
4. 文書から答えが得られる想定で質問を作成してください
5. 客観的な事実を求める質問を作成してください
6. 文書を読まないと回答できない質問を作成してください
7. 固有名詞は避けて作成してください
8. 流暢で自然な日本語で作成してください
9. 状況から上記の条件を満たす質問が作成できない場合は、'None'とだけ答えてください

Template prompts of scoring evaluation (original)

文書に対して質問応答を行うモデルの回答を評価してください。

評価の対象:
評価対象として、以下の3つ組が与えられます。
- 質問: 文書の内容に関する質問
- 正解: 文書に基づいた、正確かつ必要十分な回答
- 回答: 評価対象のモデルが生成した回答

回答の条件:
評価対象の回答は、以下の条件で生成されたものである。
1. 回答は、文書に書かれている客観的な事実のみをもとに生成すること。
2. 回答は、日本語の文章で生成すること。
3. 文書の内容から回答が生成できない場合は、回答できない理由を生成すること。

スコアの基準:
質問の文脈において、回答がどれだけ正解と意味的に一致しているかを基準とし、以下の5段階で採点してください。
5: 優秀 - 正解と完全に一致している。
4: 良い - 正解と回答に軽微な違いがある。
3: 普通 - 正解と回答に明確な違いがある。
2: 悪い - 正解と回答は大きく異なっている。
1: 非常に悪い - 正解と解答が完全に異なっているまたは、回答が日本語を主に書かれてない場合や、正解として回答できない理由が必要であるにもかかわらず、その理由が回答に明記されていない場合。

出力形式:
- スコア（1〜5 の整数）のみを出力してください。
- 説明は一切追加しないこと。

質問: Question
正解: Answer
回答: Prediction
スコア: